USER GUIDE
LFS TWO-QUARTER AND FIVE-QUARTER LONGITUDINAL DATASETS

Introduction

1. The Labour Force Survey (LFS) is a household survey, gathering information on a wide range of labour force characteristics and related topics. Since 1992 it has been conducted on a quarterly basis, with each sample household retained for five consecutive quarters, and a fifth of the sample replaced each quarter. The survey was designed to produce cross-sectional data, but in recent years it has been recognised that linking together data on each individual across quarters would produce a rich source of longitudinal data.

2. There are however methodological problems which could distort the data resulting from this linking. These fall into two main groups: biases arising from non-response and the sample attrition arising from it; and biases arising from response errors, particularly their effects in producing spurious flows between economic activity states. ONS has therefore undertaken a joint research project with Southampton University to address these methodological issues. This project has now produced a satisfactory methodology for compensating for the biasing effects of non-response, and a procedure has been developed for applying it in longitudinal datasets linking two or five adjacent quarters.

3. This guide describes the two-quarter and five-quarter longitudinal LFS datasets and how to use them. It describes briefly how they are produced, but does not give details of the methodological development - this is covered in detail in paper 17 of the GSS Methods and Quality series, entitled “Methodological Issues in the Production and Analysis of Longitudinal Data from the Labour Force Survey” by Paul Clarke and Pam Tate.

4. It is also important to note that the second methodological problem, that of response error bias, is still under investigation. Progress so far on this is also described in the Methods and Quality paper. This guide includes some discussion of the possible effects of response error bias on analyses of the longitudinal datasets, including which kinds of analyses are more or less likely to be affected, in the light of our present state of knowledge from the investigation. This guidance is highly provisional, and will be revised as the investigation proceeds.

Datasets

5. The quarterly LFS started in spring 1992, but the rotational pattern of the sample was not established until winter 92/93, therefore this is the first quarter available for longitudinal linking. Two-quarter longitudinal datasets have been produced for all pairs of adjacent quarters from winter 1992/93 onwards - for example, the winter 1992/93 dataset was linked with the spring 1993 dataset. Five-quarter longitudinal datasets have also been produced for the same period, for example linking spring 1993 with spring 1994 and containing data from all five waves of the survey. New datasets will continue to be created as further LFS quarterly

data becomes available. All the datasets are available as portable SPSS files, (those with .por extension), which are straightforward to access in either SPSS or SAS.

Coverage

6. Since the focus of analyses of these datasets was expected to be the population of working age, it was decided, in consultation with major customers, to restrict the datasets to women aged 15 to 59 at the first quarter and men aged 15 to 64 at the first quarter.

7. For the period from winter 1995/6, the datasets cover the UK. The Northern Ireland survey did not change from an annual to a quarterly survey until winter 1994/95, and the rotation pattern of the sample was not fully established until winter 1995/96, therefore the longitudinal datasets which include quarters before winter 1995/96 cover just Great Britain.

8. The small proportion of people in the sample whose data, at any of the linked quarters, had been imputed by rolling forward from the previous interview, were excluded from the longitudinal datasets.

9. From spring 1996 onwards, with the introduction of the household matrix approach to gathering data on the people present in the household, a small proportion of people in the sample have no data available on economic activity. People with no data on economic activity at one or more of the linked quarters have been excluded from the longitudinal datasets.

Linking procedure

10. The linked datasets are created as follows. The regular quarterly individual-level LFS dataset for the first of the quarters to be linked is used to produce a reduced cross-sectional dataset confined to the age range and variables to be used for the longitudinal dataset. A unique identification variable PERSID is created by combining the system variables QUOTA, WEEK, W1YR, QRTR, ADD, HHLD and RECNO. A similar procedure is followed for the other quarters. The reduced cross-sectional datasets for the two or five quarters to be linked are matched by the variable PERSID, and checked to ensure that the cases linked match also on sex and date of birth. All unmatched cases are dropped, as are all cases where the data were rolled forward, or where there are no data on economic activity, at any of the quarters.

Variables

11. Because of the resources involved in production and the size of the resultant datasets, the longitudinal datasets include only a subset of the full LFS variable set. This subset has been agreed in consultation with users and represents the most important and commonly used variables covering the main areas of the survey. A full variable list has been created which shows the content of each of the datasets. Users who have access to the quarterly datasets and who want to add in extra variables can do so by creating the PERSID variable and simply merging the variables on – ONS can supply SPSS code which enables users to do this.

12. When the linked datasets are created, all the variables relating to the first of the linked quarters are renamed, with a suffix of 1 added to the original variable name, and all the variables relating to the second of the linked quarters have a suffix of 2 added to the original variable name, and so on. For example, if we link together the summer 97 and summer 98 quarters, then the variable TEN96 from the first quarter, summer 97, becomes TEN961 in the linked dataset, and TEN96 from the second quarter, autumn 97, becomes TEN962, and so on until the summer 98 variable becomes TEN965. This is true for all the variables in all the datasets, except for the system variables used to create the identification variable PERSID, the variable PERSID itself, and the variables for sex and date of birth which are used for checking that the match between the two quarters is correct. These must be identical for each of the quarters being linked and therefore have no suffix.

13. Some of the variables are not available for all quarterly interviews and are therefore not available in one or both of the quarters of some of the linked datasets. For example, the variable HITQUA is only available for Spring (MM) and Autumn (SN) quarters from 1996. Therefore it will be available for the first quarter of the spring summer 2000 linked dataset (as HITQUA1) but not the second quarter. Similarly, it will be available for the second quarter of the winter 99 spring 2000 dataset (as HITQUA2) but not available for the first quarter.

14. As SPSS only allows variable names to be eight characters long, a few variable names which are already eight characters long have to be amended when the suffix is added. These are as follows:

	1st Qtr	**2nd Qtr (3^{rd}, 4^{th}, 5^{th})**
NMANAGE2	NMNAGE21	NMNAGE22/3/4/5
IOUTCOME	IOUTCOM1	IOUTCOM2/3/4/5
SHFTWK99	SHFTWK91	SHFTWK92/3/4/5

15. A variable FLOW has been added to the datasets. It gives in a convenient form the categories relating to labour force gross flows, distinguishing between states in and outside working age. The codes and categories are as follows:

1	Aged 15 at both quarters	
2	Entrant to working-age between first and final quarter	
3	In employment at first quarter; in employment at final quarter	(EE)
4	In employment at first quarter; unemployed at final quarter	(EU)
5	In employment at first quarter; inactive at final quarter	(EN)
6	Unemployed at first quarter; in employment at final quarter	(UE)
7	Unemployed at first quarter; unemployed at final quarter	(UU)
8	Unemployed at first quarter; inactive at final quarter	(UN)
9	Inactive at first quarter; in employment at final quarter	(NE)
10	Inactive at first quarter; unemployed at final quarter	(NU)
11	Inactive at first quarter; inactive at final quarter	(NN)
12	Reached retirement age by final quarter	

For the two-quarter datasets this variable shows the flow over a 3 month period, while for the five-quarter datasets it shows the flow over a 12 month period.

16. In addition, a variable ANFLOW has been added to the five-quarter datasets. It gives categories relating to labour force gross flows across all five of the linked quarters. There are 243 possible sequences over five quarters, many of which will have very small frequencies, particularly the ones involving 3 or 4 moves. For this reason a simplified categorisation is presented which combines together those sequences where only the timing differs - for example, all cases which start in employment and end in unemployment (with no other transitions) are in category 4 below, regardless of the wave in which they became unemployed. The codes and categories are as follows:

1	In employment in all quarters	(E)
2	Unemployed in all quarters	(U)
3	Inactive in all quarters	(N)
4	In employment at first quarter; unemployed at final quarter	(EU)
5	In employment at first quarter; inactive at final quarter	(EN)
6	Unemployed at first quarter; inactive at final quarter	(UN)
7	Unemployed at first quarter; in employment at final quarter	(UE)
8	Inactive at first quarter; in employment at final quarter	(NE)
9	Inactive at first quarter; unemployed at final quarter	(NU)
10	Employed at first; unemployed; in employment at final quarter	(EUE)
11	Employed at first; inactive; in employment at final quarter	(ENE)
12	Unemployed at first; inactive; unemployed at final quarter	(UNU)
13	Unemployed at first; employed; unemployed at final quarter	(UEU)
14	Inactive at first; employed; inactive at final quarter	(NEN)
15	Inactive at first; unemployed; inactive at last quarter	(NUN)
16	Employed at first; unemployed; inactive at final quarter	(EUN)
17	Employed at first; inactive; unemployed at final quarter	(ENU)
18	Unemployed at first; employed; inactive at final quarter	(UEN)
19	Unemployed at first; inactive; employed at final quarter	(UNE)
20	Inactive at first; employed; unemployed at final quarter	(NEU)
21	Inactive at first; unemployed; employed at final quarter	(NUE)
22	3 or 4 moves between categories	

Weighting

17. The weighting factors (variable name LGWT) for these datasets serve two purposes. They compensate for non-response bias, and also produce estimates at the level of the population. The calculation of weighting factors for the two-quarter datasets involves the following stages:

(i) Initial prior weights are calculated such that they reproduce the distribution of the cross-sectional sample from the first quarter according to the tenure/landlord categories: owned; rented from local authority/housing association; privately rented.

(ii) These prior weights are then multiplied by a single grossing factor, (except for Northern Ireland where this factor is again multiplied by an adjustment factor to compensate for the different sampling fraction), such that the weighted sample cases sum to the overall population control total (described below). This results in the prior weights used in the calculation of the final weights (described below).

(iii) A process of calibration weighting (also known as generalised raking) is then applied to the sample, using CALMAR software (see Elliot 1997). This process minimises the distance between the prior and final weights, while constraining the final weights simultaneously to several marginal distributions or control totals. Four sets of control totals are used:

(a) the population estimates used for weighting the second quarter's cross-sectional LFS dataset, for the selected age range, classified by sex and age (in single years to 24, then five-year age groups) - this produces estimates as close as possible to the population available for sampling in both the linked quarters;

(b) the population estimates used for weighting the second quarter's cross-sectional LFS dataset, for the selected age range, classified by region;

(c) the weighted cross-sectional estimates from the second linked quarter for the selected age range classified by broad economic activity categories: in employment; unemployed; economically inactive;

(d) the weighted cross-sectional estimates from the first linked quarter for the selected age range classified by broad economic activity categories: in employment; unemployed; economically inactive, adjusted to the same total as (a) to (c) by reducing the economically inactive category as necessary.

CALMAR is run using the logit method, with the ratio of the final to prior weights constrained to the range 0.5 to 1.7.

The elements (iii) (a) and (b) in this process produce population-level estimates, and also contribute to some extent to compensating for non-response bias; the other elements complete the compensation for non-response bias.

The extension of this method to create five-quarter datasets consists of constraining to the cross-sectional economic activity distribution at each of the five quarters. This involves repeating the constraint in (iii)d for the second, third and fourth quarters as well as the first, adjusting in the same way to achieve a total consistent with the fifth quarter. When running CALMAR to create five-quarter datasets, wider limits have to be set for the ratio of final to prior weights, typically 0.6 to 2.3.

Sample sizes and threshold levels

18. Because of sampling variability, the smaller the group being estimated the poorer the precision of the estimate becomes, until eventually the estimate is not reliable enough to be used. (See Volume 1 of the LFS User Guide for a detailed discussion.) For the regular quarterly cross-sectional LFS datasets, a publication threshold is set at 10,000 (i.e. estimates below 10,000 are not published), at which level the standard error is about 20% of the estimate, and the 95% confidence interval for the estimate is about +/-4,000. For the two-quarter longitudinal datasets, the same principle applies, but the number of sample cases

available for linkage is smaller (usually around 60,000), so the threshold level for these datasets is 17,000.

19. Because of the lower number of cases available for linking and higher attrition, the five-quarter datasets contain only around 11,000 cases each compared to 60,000 on the two-quarter datasets. Therefore the results are subject to greater variability and the threshold levels for producing reliable estimates are much higher than the 17,000 recommended for the two-quarter datasets. However, it is possible to combine results from several datasets to reduce the threshold level. The table below shows the thresholds that should be used for different categories of ANFLOW:

Number of datasets used	1	2	3	4
Category 1 or 3 (E or N)	68,000	34,000	23,000	17,000
Category 2 (U)	130,000	65,000	44,000	33,000
All other categories	100,000	50,000	33,000	25,000

20. For estimates below these figures, the standard error is likely to be greater than 20% of the estimate and therefore the estimate should not be used. The figures are higher for the unemployed category because of the design effect and higher attrition within this group. For some of the other categories, particularly those involving more than one transition, there may be very few cases present in each dataset. Therefore it may be necessary to combine categories or use several datasets to get a reliable result.

Some points on longitudinal analysis, including the implications of response error bias

21. All analyses should be run weighted by LGWT, otherwise the results will be distorted by non-response bias, and possibly misleading.

22. Careful thought is needed about the precise coverage of any analysis – is it the population of working age at the first quarter, the second quarter, or both quarters? The variable FLOW can be used to select any of these groups: codes 3 to 12 give working age at the first quarter, 2 to 11 at the second quarter, and 3 to 11 at both quarters.

23. Most analyses of interest are likely to be cross-tabulations of a characteristic at the first quarter with a characteristic at the second or fifth quarter, often restricted to a subgroup. Some examples are: lone parents of working age at both quarters by sex and age of youngest child and by economic activity at both quarters; young people aged 18 to 24 unemployed at the first quarter by educational qualification and economic activity at the last quarter; people reaching retirement age by the last quarter by economic activity at both quarters and by reason for inactivity if inactive. Doing analyses of this kind, the numbers of cases in some cells can very quickly fall below the threshold level.

24. Research so far on response error has been based on empirical analysis of differences in levels of transitions between different economic activity categories and of apparent internal inconsistencies. The findings so far are therefore provisional, and will be updated with the results of further research. However, the initial investigations have provided evidence

supporting the suggestion that response error is likely to affect the longitudinal datasets, probably in the direction of an upward bias in estimates of gross flows between different broad economic activity categories. It has also provided some tentative indications of transitions and subgroups particularly likely to be affected. These are transitions between unemployment and inactivity, transitions between part-time employment and either unemployment or inactivity, for women any transitions involving unemployment, and for students transitions between employment and unemployment. However, some of the apparent inconsistencies may be caused by genuine volatility (repeated movements back and forth between different economic activity states) rather than by response error.

Contact details

All enquiries about the longitudinal datasets should be directed to Mike Young, ONS, Room B4/03, 1 Drummond Gate, London SW1V 2QQ, telephone 020 7533 6160, e-mail mike.young@ons.gov.uk.

Reference

Elliot, D (1997) *Software to weight and gross survey data.* GSS Methodology Series No 1.

SPSS Code For Adding Variables To Five Quarter Longitudinal Datasets

Introduction

Due to the resources involved in production and the size of the resultant datasets the longitudinal datasets only include a subset of the full LFS variable set. However if users have access to the quarterly LFS datasets they can add extra variables by using the following code. The code works by creating the PERSID variable in the quarterly datasets and merging the required variables onto the longitudinal dataset.

Example of the SPSS code

Below is an example of the SPSS code for adding the extra variables PUBLIC and MAINME to the Summer 2000-01 (lgsum001.sav) 5 quarter longitudinal dataset.

Since SPSS only allows variable name lengths to be eight characters long, when adding variables with names of eight characters one of the characters will have to be dropped so the quarter suffix can be added. See Longitudinal User Guide sections 12 and 14.

Please note that the parts highlighted in red must be checked each time the code is run to make sure that the file names, weights and variable names are all correct.

*** Code for adding variables to 5Q Longitudinal Datasets.**

```
GET FILE='D:\ja00.sav'
/KEEP
AGE SEX QUOTA WEEK W1YR QRTR ADD HHLD RECNO INTWT
PUBLIC MAINME
/RENAME=(PUBLIC MAINME=PUBLIC1 MAINME1).
```

```
*  Select persons aged 15 to 59/64 with non-zero weights.
SELECT IF (AGE >=15 AND ((AGE <= 64 & SEX=1) OR (AGE<=59 & SEX=2))).
SELECT IF INTWT >0.

* Create a person serial number.

COMPUTE
PERSID=(QUOTA*10000000000)+(WEEK*100000000)+(W1YR*10000000)
+(QRTR*1000000)+(ADD*10000)+(HHLD*100)+RECNO.

* Sort the cases in preparation for linking.

SORT CASES BY PERSID.

*Name Q1 outfile.

SAVE OUTFILE='C:\ja00cs.sav'
/RENAME=(AGE=AGE1).

GET FILE='D:\sn00.sav'
/KEEP
AGE SEX QUOTA WEEK W1YR QRTR ADD HHLD RECNO INTWT
PUBLIC MAINME
/RENAME=(PUBLIC MAINME=PUBLIC2 MAINME2).


*  Select persons aged 15 to 60/65 with non-zero weights.
SELECT IF (AGE >=15 AND ((AGE <= 65 & SEX=1) OR (AGE<=60 & SEX=2))).
SELECT IF INTWT >0.


* Create a person serial number.

COMPUTE
PERSID=(QUOTA*10000000000)+(WEEK*100000000)+(W1YR*10000000)
+(QRTR*1000000)+(ADD*10000)+(HHLD*100)+RECNO.

* Sort the cases in preparation for linking.

SORT CASES BY PERSID.

*Name Q2 outfile.

SAVE OUTFILE='C:\sn00cs.sav'
/RENAME=(AGE=AGE2).
```

```
GET FILE='D:\d00f.sav'
/KEEP
AGE SEX QUOTA WEEK W1YR QRTR ADD HHLD RECNO INTWT
PUBLIC MAINME
/RENAME=(PUBLIC MAINME=PUBLIC3 MAINME3).



*  Select persons aged 15 to 59/64 with non-zero weights.
SELECT IF (AGE >=15 AND ((AGE <= 65 & SEX=1) OR (AGE<=60 & SEX=2))).
SELECT IF INTWT >0.

* Create a person serial number.

COMPUTE
PERSID=(QUOTA*10000000000)+(WEEK*100000000)+(W1YR*10000000)
+(QRTR*1000000)+(ADD*10000)+(HHLD*100)+RECNO.

* Sort the cases in preparation for linking.

SORT CASES BY PERSID.

*Name Q3 outfile.

SAVE OUTFILE='C:\d00fcs.sav'
/RENAME=(AGE=AGE3).


GET FILE='D:\mm01.sav'
/KEEP
AGE SEX QUOTA WEEK W1YR QRTR ADD HHLD RECNO INTWT
PUBLIC MAINME
/RENAME=(PUBLIC MAINME=PUBLIC4 MAINME4).



*  Select persons aged 15 to 59/64 with non-zero weights.
SELECT IF (AGE >=15 AND ((AGE <= 65 & SEX=1) OR (AGE<=60 & SEX=2))).
SELECT IF INTWT >0.

* Create a person serial number.

COMPUTE
PERSID=(QUOTA*10000000000)+(WEEK*100000000)+(W1YR*10000000)
+(QRTR*1000000)+(ADD*10000)+(HHLD*100)+RECNO.
```

```
* Sort the cases in preparation for linking.

SORT CASES BY PERSID.

*Name Q4 outfile.

SAVE OUTFILE='C:\mm01cs.sav'
/RENAME=(AGE=AGE4).

GET FILE='D:\ja01.sav'
/KEEP
AGE SEX QUOTA WEEK W1YR QRTR ADD HHLD RECNO INTWT
PUBLIC MAINME
/RENAME=(PUBLIC MAINME=PUBLIC5 MAINME5).



*  Select persons aged 15 to 59/64 with non-zero weights.
SELECT IF (AGE >15 AND ((AGE <= 65 & SEX=1) OR (AGE<=60 & SEX=2))).
SELECT IF INTWT >0.

* Create a person serial number.

COMPUTE
PERSID=(QUOTA*10000000000)+(WEEK*100000000)+(W1YR*10000000)
+(QRTR*1000000)+(ADD*10000)+(HHLD*100)+RECNO.

* Sort the cases in preparation for linking.

SORT CASES BY PERSID.

*Name Q5 outfile.

SAVE OUTFILE='C:\ja01cs.sav'
/RENAME=(AGE=AGE5).


MATCH FILES
FILE='C:\ja00cs.sav'
/FILE='C:\sn00cs.sav'
/FILE='C:\d00fcs.sav'
/FILE='C:\mm01cs.sav'
/FILE='C:\ja01cs.sav'
/FILE='D:\lgsum001.sav
/BY PERSID
```

```
/DROP=intwt.

* Select cases with information for both quarters that were not imputed
* in either quarter and drops unlinked cases.

SELECT IF NOT SYSMIS (IOUTCOM1).

*Name new 5 quarter dataset with added variables.

SAVE OUTFILE='c:\lgsum001new.sav'.
```

Contact details

All enquiries about the longitudinal datasets should be directed to Jeremy Reuben, ONS, Room B4/04, 1 Drummond Gate, London SW1V 2QQ, telephone 020 7533 6320, e-mail jeremy.reuben@ons.gov.uk..